Hi-kvadrat test (χ^2)

Testiranje veze između kategorijskih varijabli — za dva časa

Predavač

Srednja škola

2026.

Plan predavanja — dva časa

Čas 1: Zašto i šta?

1. Kategorijske varijable i dvosmjerne tablice
2. Šta znači "veza" između varijabli?
3. Očekivani brojevi — kako ih izračunamo?
4. Hi-kvadrat statistika — intuicija i formula

Čas 2: Kako i kada?

1. Hi-kvadrat distribucija i stepeni slobode
2. P-vrijednost i zaključak
3. Kad smijemo koristiti test?
4. Simpsonov paradoks — oprez!
5. Praktični primjeri i zadaci

Ključno pitanje oba časa

Postoji li stvarna veza između dvije kategorijske varijable — ili su razlike koje vidimo samo slučajnost?

Čas 1 — Podsjetnik: kategorijske varijable

Kategorijska varijabla

Varijabla koja svrstava pojedinca u jednu od **nekoliko grupa** (kategorija). Nema smislene numeričke vrijednosti.

Primjeri kategorijskih varijabli

- **Pol**: muški / ženski
- **Omiljeni predmet**: matematika / biologija / istorija...
- **Tip prijevoza**: pješice / autobus / auto / bicikl
- **Godišnje doba**: proljeće / ljeto / jesen / zima
- **Krvna grupa**: A / B / AB / 0

Ključno pitanje ovog časa

Ako imamo **dvije** kategorijske varijable — npr. pol i omiljeni predmet — postoji li između njih veza?

Da li djevojke i dječaci imaju *različite* preferencije prema predmetima?

Kako to **statistički** testiramo?

Dvosmjerna tablica — osnova svega

Dvosmjerna (kontingencijska) tablica

Prikazuje **frekvencije** (broj slučajeva) za sve kombinacije dviju kategorijskih varijabli.

Primjer: Omiljeni predmet i pol (izmišljeni podaci, $n = 200$)

	Matematika	**Biologija**	**Istorija**	**UKUPNO**
Muški	48	26	26	**100**
Ženski	32	44	24	**100**
UKUPNO	**80**	**70**	**50**	**200**

Šta tablice govore?

- **Marginalne frekvencije**: zbroji redova i kolona
- **Unutarnje ćelije**: broj u svakoj kombinaciji

Pitanje

Gledamo li u ove podatke, čini se da dječaci više vole matematiku, a djevojke biologiju. Ali... je li to *stvarna razlika* ili samo slučajnost uzorka?

Procenti pomažu — ali samo do neke mjere

Korak 1: Izračunaj procente unutar svake grupe (reda)

Postotak opisuje distribuciju *jedne* varijable unutar svake kategorije *druge*.

Proc. po polu

	Matematika	**Biologija**	**Istorija**	**Ukupno**
Muški	$48/100 = 48\%$	$26/100 = 26\%$	$26/100 = 26\%$	100%
Ženski	$32/100 = 32\%$	$44/100 = 44\%$	$24/100 = 24\%$	100%

Šta vidimo?

- Dječaci: **48%** voli matematiku
- Djevojke: samo **32%** voli matematiku
- Djevojke: **44%** voli biologiju
- Dječaci: samo **26%** voli biologiju

Problem postotaka

Proc. opisuju *uzorak* koji imamo. Ali ne možemo samo pogledati i reći: "Aha, veza postoji!". Možda je ova razlika **slučajnost** - drugi uzorak iz iste populacije mogao bi dati sasvim drugačije proc. Trebamo **stat. test**.

Nulta hipoteza: "Nema veze"

Osnovna ideja testiranja

Kao i svaki statistički test, hi-kvadrat test počinje s **nultom hipotezom**:

H_0: **Nema veze između dvije varijable u populaciji.**

Šta znači "nema veze"?

Ako nema veze između pola i omiljenog predmeta, onda:

- Distribucija omiljenog predmeta je **ista** za dječake i djevojke
- Znanje da je neko dječak ili djevojka nam **ništa ne govori** o omiljenom predmetu
- Varijable su **nezavisne** jedna od druge

Logika testiranja

1. Pretpostavi H_0 (nema veze)
2. Izračunaj **što bismo očekivali** vidjeti u tablici ako H_0 vrijedi
3. Usporedi *opažene* s *očekivanim* brojevima
4. Ako su previše različiti - H_0 je **nevjerovatna**

Ključni korak

Kako izračunati očekivane brojeve? To je srž

Očekivani brojevi — korak po korak

Formula za očekivani broj u svakoj ćeliji

$$E = \frac{\text{ukupno reda} \times \text{ukupno kolone}}{\text{ukupno tablica}}$$

Primjer: Ćelija "Muški + Matematika"

- Ukupno muških: 100
- Ukupno koji vole matematiku: 80
- Ukupan broj ispitanika: 200

$$E(\text{Muški, Mat.}) = \frac{100 \times 80}{200} = \frac{8000}{200} = 40$$

Intuicija: ako 80/200 = 40% svih voli matematiku, i nema razlike po polu, onda bi i od 100 dječaka trebalo 40% = 40 voljeti matematiku.

Svih 6 očekivanih vrijednosti — provjera

Primjena formule $E = \text{red} \times \text{kolona}/n$ na sve ćelije

	Mat. (80 ukupno)	**Bio.** (70 ukupno)	**Ist.** (50 ukupno)
Muški (100)	$100 \times 80/200 = 40$	$100 \times 70/200 = 35$	$100 \times 50/200 = 25$
Ženski (100)	$100 \times 80/200 = 40$	$100 \times 70/200 = 35$	$100 \times 50/200 = 25$

Uočite!

Kad nema veze (H_0), *oba* reda imaju **iste** očekivane vrijednosti jer su marginalni zbroji redova jednaki (100 i 100).

Provjera: Zbroj reda E: $40 + 35 + 25 = 100$ ✓
Zbroj kolone E: $40 + 40 = 80$ ✓

Što nam ovo govori?

Ako **nema razlike** između muških i ženskih u preferenciji predmeta, onda bismo u svakoj ćeliji trebali vidjeti upravo ove brojeve.

Sve što se razlikuje od ovih E vrijednosti — razlika je od H_0.

Na sljedećem slajdu mjerimo koliko su O i E

Hi-kvadrat statistika — intuicija

Osnovna ideja

Mjerimo **koliko su opaženi (O) brojevi daleko od očekivanih (E)**. Što je veća razlika — to je jači dokaz da H_0 nije tačna.

Zašto ne koristimo samo $O - E$?

- $O - E$ može biti negativan — zbrojevi bi se poništili
- Razlika od 5 je *važna* ako je $E = 10$, ali *beznačajna* ako je $E = 1000$
- Zato koristimo $(O - E)^2/E$ — kvadriramo (sve pozitivno) i podijelimo s E (relativna veličina)

Formula hi-kvadrat statistike

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

Suma ide **po svim ćelijama** tablice.

- O = opaženi broj (iz podataka)
- E = očekivani broj (iz H_0)
- $\chi^2 \geq 0$ uvijek
- Veći $\chi^2 \Rightarrow$ jači dokaz protiv H_0

Izračun χ^2 — naš primjer korak po korak

Opaženi vs. Očekivani brojevi

	O (Mat.)	E (Mat.)	O (Bio.)	E (Bio.)	O (Ist.)	E (Ist.)
Muški	48	40	26	35	26	25
Ženski	32	40	44	35	24	25

Doprinos svake ćelije χ^2

Ćelija	Izračun	Doprinos
Muški + Mat.	$(48-40)^2/40 = 64/40$	1,60
Ženski + Mat.	$(32-40)^2/40 = 64/40$	1,60
Muški + Bio.	$(26-35)^2/35 = 81/35$	2,31
Ženski + Bio.	$(44-35)^2/35 = 81/35$	2,31
Muški + Ist.	$(26-25)^2/25 = 1/25$	0,04
Ženski + Ist.	$(24-25)^2/25 = 1/25$	0,04

Kraj prvog časa

Podsjetnik:

Izračunali smo $\chi^2 = 7{,}90$ za naš primjer.

Na drugom času: šta nam taj broj zapravo govori?

Čas 2 — Podsjetnik s prvog časa

Šta smo naučili

- Dvosmjerne tablice prikazuju veze između **dviju kategorijskih varijabli**
- H_0: **Nema veze** između varijabli
- Očekivane vrijednosti: $E = \frac{\text{red} \times \text{kolona}}{n}$
- Hi-kvadrat statistika: $\chi^2 = \sum \frac{(O-E)^2}{E}$
- Dobili smo $\chi^2 = 7{,}90$ za naš primjer

Pitanje koje sad rješavamo

Koliko velik mora biti χ^2 da bismo odbacili H_0?

Odgovor ovisi o:

1. **Stepeni slobode** — veličina tablice
2. **Razini značajnosti** — koliko smo strogi ($\alpha = 0{,}05$?)

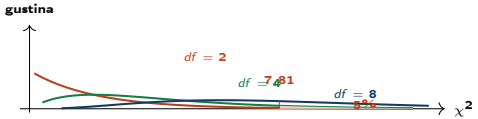
Trebamo razumjeti **hi-kvadrat distribuciju**.

Hi-kvadrat distribucija — šta izgleda?

Hi-kvadrat distribucija

Kad je H_0 tačna, χ^2 statistika prati **hi-kvadrat distribuciju**. Ova distribucija:

- Uzima samo **nenegativne** vrijednosti ($\chi^2 \geq 0$)
- Nagnuta je **udesno** (iskošena pozitivno)
- Ovisi o parametru zvanom **stepeni slobode (df)**



Veći df ⇒ kriva se pomiče desno.

Zašto udesno nagnuta?

Kad je H_0 tačna, χ^2 je **mali**. Veliki χ^2 je **rijedak** i ukazuje da H_0 nije tačna.

Analogija: 9 glava od 10 bacanja kovanice — sumnjamo na kovanicu!

P-vrijednost

P-vrijednost = **površina desnog repa** desno od naše χ^2.

Stepeni slobode — zašto su važni?

Formula za stepene slobode

Za tablicu s r redova i c kolona:

$$df = (r-1)(c-1)$$

Intuicija

Kad znamo marginalne zbroje tablice, koliko ćelija možemo **slobodno popuniti** prije nego što su preostale određene?

Tablica 2×3: znamo zbroje redova i kolona. Slobodnih ćelija: $(2-1)(3-1) = 2$.

Zamislite: popunite samo gornji lijevi 1×2 *kvadrant. Ostalo je automatski određeno!*

Naš primjer (2×3 tablica)

$df = (2-1)(3-1) = 1 \times 2 = 2$

?	**?**	*auto*	100
auto	*auto*	*auto*	100
80	70	50	200

Samo 2 ćelije su slobodne (označene **?**). Ostatak je određen zbrojevima.

Kritične vrijednosti — tablica za brzo zaključivanje

Tablica kritičnih vrijednosti χ^2

Odbacujemo H_0 ako je naša χ^2 **veća** od kritične vrijednosti.

Stepeni slobode (df)	**P < 0,10**	**P < 0,05**	**P < 0,01**
1	2,71	3,84	6,63
2	4,61	**5,99**	9,21
3	6,25	7,81	11,34
4	7,78	9,49	13,28
5	9,24	11,07	15,09
6	10,64	12,59	16,81

Naš primjer: $df = 2$, $\chi^2 = 7{,}90$

Pogledamo red $df = 2$: Kritična vrijednost za $P < 0{,}05$: **5,99**. Kritična vrijednost za $P < 0{,}01$:

Zaključak

Postoji **statistički značajna veza** između pola i omiljenog predmeta u populaciji ($P < 0{,}05$).

Pravi primjer: Digitalna pismenost i tačno prepoznavanje vijesti

Istraživanje

Ispitanici su čitali 5 tvrdnji i trebali prepoznati koje su *činjenice*, a koje su *mišljenja*. Svrstani su po razini "digitalne pismenosti". Zabilježeno: je li svaki ispitanik ispravno identificirao **svih 5 činjenica**.

Dvosmjerna tablica ($n = 1000$)

Digitalna pismenost	**Sve 5 tačne**	**Nije sve 5**	**Ukupno**
Visoka	168	312	**480**
Srednja	70	280	**350**
Niska	22	148	**170**
Ukupno	**260**	**740**	**1000**

Postotci po grupi

Visoka pismenost: $168/480 = 35{,}0\%$ tačnih; Srednja: $70/350 = 20{,}0\%$ tačnih; Niska: $22/170 = 12{,}9\%$

Pitanje

Razlika (35% vs. 13%). To je *stvarna razlika* u popul. ili slučajnost uzorka? Primijenimo hi-kvadrat test!

Pravi primjer: Izračun χ^2 korak po korak

Korak 1: Očekivane vrijednosti

	Sve 5 tačne (E)	**Nije sve 5** (E)
Visoka	$480 \times 260/1000 = 124{,}8$	$480 \times 740/1000 = 355{,}2$
Srednja	$350 \times 260/1000 = 91{,}0$	$350 \times 740/1000 = 259{,}0$
Niska	$170 \times 260/1000 = 44{,}2$	$170 \times 740/1000 = 125{,}8$

Korak 2: Doprinos svake ćelije

	$(O-E)^2/E$ (tačne)	$(O-E)^2/E$ (nije)	Zbir po redu
Visoka	$(168-124{,}8)^2/124{,}8 = 14{,}95$	$(312-355{,}2)^2/355{,}2 = 5{,}25$	20,20
Srednja	$(70-91{,}0)^2/91{,}0 = 4{,}85$	$(280-259{,}0)^2/259{,}0 = 1{,}70$	6,55
Niska	$(22-44{,}2)^2/44{,}2 = 11{,}15$	$(148-125{,}8)^2/125{,}8 = 3{,}92$	15,07
Ukupno χ^2			41,82

Pravi primjer: Zaključak

Stepeni slobode

Tablica 3×2: $df = (3-1)(2-1) = 2$

Poređenje s tablicom ($df = 2$)

- Krit. vrijednost za $P < 0{,}05$: **5,99**
- Krit. vrijednost za $P < 0{,}01$: **9,21**

Naša $\chi^2 = 41{,}82$

$41{,}82 \gg 9{,}21 \Rightarrow P < 0{,}01$

Odbacujemo H_0 s velikom sigurnošću.

Zaključak

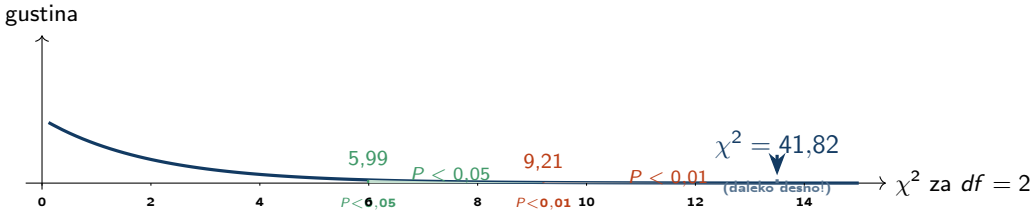
Postoji **statistički visoko značajna veza** između razine digitalne pismenosti i sposobnosti prepoznavanja činjeničnih tvrdnji ($P < 0{,}01$).

Digitalno pismeniji **statistički značajno bolje** prepoznaju vijesti.

Vizualizacija: $\chi^2 = 41{,}82$ na hi-kvadrat distribuciji

Hi-kvadrat distribucija za $df = 2$ — gdje pada naša vrijednost?

Što je χ^2 dalje u desnom repu, to je P-vrijednost manja i dokaz jači.



Zaključak: $\chi^2 = 41{,}82 \gg 9{,}21$ — P-vrijednost je mnogo manja od 0,01

Kad smijemo koristiti hi-kvadrat test?

Uvjeti za valjanu primjenu

1. **Slučajni uzorak (SRS)**: podaci moraju biti prikupljeni slučajno
2. **Dovoljna veličina uzorka**: svi očekivani brojevi moraju biti dovoljno veliki

Pravilo o veličini očekivanih vrijednosti

Hi-kvadrat test je **valjan** ako:

- **Svi** pojedinačni $E \geq 1$
- Najviše **20%** ćelija ima $E < 5$

Ako ovo nije ispunjeno — test nije pouzdan! Rješenje: spojiti male kategorije ili uzeti veći uzorak.

Provjera za naš primjer

Najmanji $E = 44{,}2$ (ćelija Niska/Tačno).
Svi $E > 5$ ✓
Niti jedna ćelija nema $E < 5$ ✓

Test je valjan!

Da je neka $E = 2{,}3$, trebali bismo spojiti kategorije "Srednja" i "Niska" pismenost.

Još jedna napomena

Hi-kvadrat test otkriva **postoji li veza**, ali ne govori **koliko je jaka** niti **u kom smjeru**. Za jačinu veze koristimo druge mjere (Cramerov V i sl.).

Simpsonov paradoks — opasna zamka!

Simpsonov paradoks

Veza koja postoji unutar svake podgrupe može nestati ili čak promijeniti smjer kada se podaci kombiniraju u jednu grupu.

Klasičan primjer: Prijem na medicinski fakultet (izmišljeni podaci)

	Ukupno		**Smjer A**		**Smjer B**	
	Primljeni	%	Primljeni	%	Primljeni	%
Muškarci	35 od 100	**35%**	30 od 60	**50%**	5 od 40	**12,5%**
Žene	25 od 80	**31%**	8 od 20	**40%**	17 od 60	**28,3%**

Paradoks!

Ukupno: muškarci imaju **bolji postotak** (35% vs. 31%).

Po smjeru: žene imaju **bolji postotak** na oba

Zašto?

Žene su se više prijavile na **Smjer B** koji je teži (niži postotak prijema za sve).

Treća varijabla (smjer) objašnjava paradoks.

Simpsonov paradoks — još jedan primjer

Baseball batting average (prosječna stopa pogodaka)

	Protiv lijevoruka bacača		**Protiv desnoruka bacača**		**Ukupno**
	Pogodak/AB	Prosjek	Pogodak/AB	Prosjek	
Igrač A	40/100	**0,400**	200/500	**0,400**	240/600 = **0,400**
Igrač B	45/100	**0,450**	250/600	**0,417**	295/700 = **0,421**

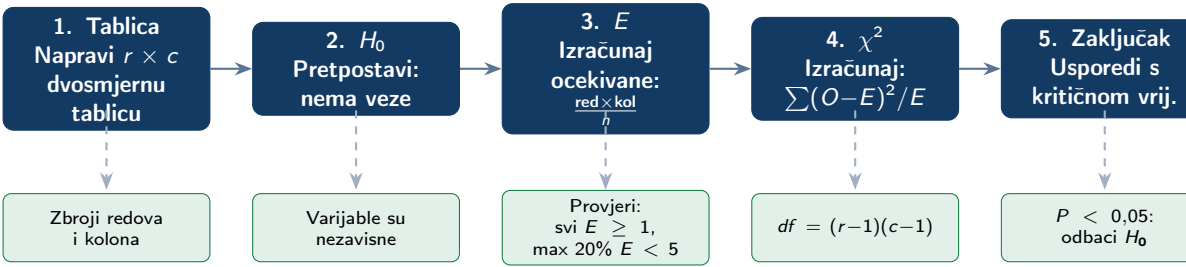
Igrač B je bolji i protiv ljevaka (0,450 vs. 0,400) i protiv dešnjaka (0,417 vs. 0,400)...

...ali Igrač A ima jednak ukupni prosjek kao Igrač B (0,400 vs. 0,421)?

Lekcija

Uvijek gledaj **podgrupu po podgrupu** kad uspoređuješ! Kombiniranje podataka može dati obmanjujuće zaključke. Hi-kvadrat test koji radi na kombiniranim podacima može propustiti ili lažno prikazati stvarne odnose.

Kompletni postupak hi-kvadrat testa



Praktični zadatak 1: Muzika i ocjene

Scenarij

Anketiramo 150 učenika o tome slušaju li muziku dok uče i njihovom prosjeku ocjena:

	Prosjek < 3	**Prosjek 3–4**	**Prosjek > 4**	**Ukupno**
Sluša muziku	22	35	18	**75**
Ne sluša	18	38	19	**75**
Ukupno	**40**	**73**	**37**	**150**

Zadaci

1. Izračunaj svih 6 očekivanih vrijednosti E
2. Izračunaj χ^2 statistiku
3. Koliki su stepeni slobode?
4. Je li χ^2 statistički značajan na $\alpha = 0{,}05$?
5. Kakav je tvoj zaključak?

Rješenja

(1) E: svi $= 75 \times 40/150 = 20$;
$75 \times 73/150 = 36{,}5$; $75 \times 37/150 = 18{,}5$

(2) $\chi^2 = \frac{(22-20)^2}{20} + \frac{(35-36{,}5)^2}{36{,}5} + \frac{(18-18{,}5)^2}{18{,}5} + \frac{(18-20)^2}{20} + \ldots = 0{,}48$

(3) $df = (2-1)(3-1) = 2$

Praktični zadatak 2: Sport i rapoloženje

Scenarij

Istraživač pita 200 tinejdžera bave li se sportom i kako se osjećaju:

	Dobro	**Srednje**	**Loše**	**Ukupno**
Bavi se sportom	85	30	5	**120**
Ne bavi se	40	30	10	**80**
Ukupno	**125**	**60**	**15**	**200**

Zadaci

1. Izračunaj sve E. Provjeri uvjet valjane primjene.
2. Izračunaj χ^2 i df.
3. Zaključi na razini $\alpha = 0{,}05$.
4. Koji bi bio Simpsonov paradoks u ovom primjeru — zamišljaj!

Rješenja

(1) E: $120\times125/200=75$; $120\times60/200=36$; $120\times15/200=9$; $80\times125/200=50$; $80\times60/200=24$; $80\times15/200=6$. *Pažnja:* $E=9$, $E=6$ — *oboje* ≥ 5 ✓

(2) $\chi^2 \approx 10{,}44$; $df = 2$

(3) $10{,}44 > 9{,}21 \Rightarrow P < 0{,}01$: **značajno!**

(4) Npr.: mlađi više sportiraju — treća varijabla.

Hvala na pažnji!

Pitanja i diskusija

"Velika razlika u uzorku nije uvijek dokaz stvarne veze —
i mala razlika u uzorku nije uvijek slučajnost."

— *osnovna ideja hi-kvadrat testa*